Developments in network location with mobile and congested facilities

Oded BERMAN
Faculty of Management, The University of Calgary, Calgary, Alberta T2N-1N4, Canada

Richard C. LARSON and Amedeo R. ODONI
Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Received June 1979
Revised January 1980

We review four facility location problems which are motivated by urban service applications and which can be thought of as extensions of the classic *Q*-median problem on networks. In problems P1 and P2 it is assumed that travel times on network links change over time in a probabilistic way. In P2 it is further assumed that the facilities (servers) are movable so that they can be relocated in response to new network travel times. Problems P3 and P4 examine the *Q*-median problem for the case when the service capacity of the facilities is finite and, consequently, some or all of the facilities can be unavailable part of the time. In P3 the facilities have stationary home locations but in P4 they have movable locations and thus can be relocated to compensate for the unavailability of the busy facilities. We summarize our main results to date on these problems.

Keywords: Facility location, network, *Q*-median problem, travel times

1. Introduction

The location of facilities is a spatial allocation of scarce resources. Cities, whose residents require a large number of diverse services, pose particularly difficult questions of facility location. Some facilities are fixed, such as libraries, schools, and out-patient clinics, whereas others are mobile service providers such as police vehicles and taxi cabs; still others have mobile service providers (or *servers*) at fixed facilities or *home locations*, such as fire engines and pumpers at fire stations and ambulances at hospitals. Some require each service request or *customer* to travel to the facility; others require the facility or a server located at a facility to travel to each customer.

When attempting to apply in cities classical deterministic location theory methods such as those derived from Q-median and/or Q-center (where Q is the number of facilities to be located) formulations, one often encounters the following problems:

(1) Travel times on urban streets are probabilistic quantities influenced by traffic conditions, weather, street repair work, traffic lights, etc.

(2) Customer arrival times and locations, as well as their associated service requirements, are also probabilistic quantities, creating various levels of spatio-temporal congestion in the system servicing these customers.

(3) Facilities do not have unlimited service capacity and thus may be *unavailable* at the time a nearby service request occurs; this, in turn, creates a pattern of workload sharing, in which facilities other than the closest facility – which may be temporarily unavailable – quite regularly service customers.

(4) Urban service systems are often multi-objective systems in which acceptable system performance is determined by a set of measures of efficiency, equity, and effectiveness. For instance, for an ambulance system a measure of efficiency is mean city-wide travel time for a given number of ambulances; a measure of equity is the magnitude of the maximum difference between neighborhood-specific mean travel times; a measure of effectiveness is the number of lives saved or hospital days reduced by actions taken by the ambulance crew.

Thus, the classical location formulations are inapplicable in complex urban settings because they assume a single-objective, deterministic world in which the closest facility is always the one utilized.

In recent years, we and our colleagues have been attempting to incorporate some of the complexities cited above into new, usually probabilistic formulations of multi-facility location problems. Our approach has been to attempt to bring into the analysis one or two such complexities at a time, recognizing that a fully comprehensive model and location algorithm remain beyond tractability at this time. Yet, it is

Table 1
Comparison of the Q-median problem with the four problems discussed in this paper

Classic Q-median problem	P1: Medians on stochastic networks	P2: Movable servers on stochastic networks	P3: Congested medians	P4: Congested movable server systems
Demand at nodes only	Demand at nodes only	Demand at nodes only	Demand at nodes only	Demand at nodes only
Unlimited service capacity	Unlimited service capacity	Unlimited service capacity	Limited service capacity [a]	Limited service capacity [a]
Fixed-location facilities	Fixed-location facilities	Movable location facilities or servers [a]	Fixed-location facilities or servers	Movable location servers [a]
Deterministic link lengths	Probabilistic link lengths [a]	Probabilistic link lengths [a]	Deterministic link lengths	Deterministic link lengths
Service by closest facility	Service by closest facility	Service by closest facility or by closest mobile server	Service by closest available facility or by closest available mobile server [a]	Service by closest available mobile server [a]

[a] Assumption different from that of the classic Q-median problem.

our hope that progress along each of several fronts will begin to indicate the types of new location results that obtain when further realism is brought to locational modeling.

In this paper, we present a review of results that have been obtained recently on a set of four problems, labelled P1 through P4 (see Table 1), which in many ways are similar to the Q-median problem on networks [8,9].

For comparative purposes we summarize the Q-median problem as follows:

(1) There are Q facilities to be located on a network $G(N, L)$.

(2) Requests for service ('customers') occur only at the nodes N.

(3) Each request is serviced by the closest facility, which is assumed to be always available to provide service.

(4) Travel time on any fraction θ of a link is assumed to be equal to (total link travel time) $\cdot\, \theta$ $(0 \leqslant \theta \leqslant 1)$.

(5) The problem is to find the set of Q facility locations that minimizes mean travel time associated with a random service request.

We have modified the Q-median problem to include such considerations as probabilistic service demands and travel times, unavailability (or congestion) of facilities and mobility of either facilities or servers located at facilities. Referring to Table 1, the problems may be summarized as follows:

P1: *Medians on stochastic networks.* This is essentially the Q-median formulation, but with the condition that travel time on each link is a discrete random variable. Our formulation, which is precisely detailed in Section 2, allows link and thus network travel times to change only at discrete time epochs and assumes that the traveller (either the customer or the server) has perfect knowledge of current network travel times. In applications, the discrete time epochs may refer to time intervals between receipt of reports of traffic conditions. Given any particular set of values for link travel times, it is assumed that each customer is serviced by the closest (in current travel times) facility. The problem is to locate the Q-facilities in order to minimize mean travel time associated with each customer, where the average is taken over all network (travel time) states. P1 has been dealt with in detail elsewhere [8,17] and thus will be treated here only briefly as a special case of P2.

P2: *Movable servers on stochastic networks.* This is the same as P1 with the additional complication that the location of each facility may be changed as network travel times change. For instance, during rush hour conditions, the city of Boston's Department of Health and Hospitals may change the location of one or more of its 'downtown' ambulances if severe traffic congestion occurs; due to a tunnel, which may become a very congested link during rush hours, one option that is currently employed is to move one of the downtown ambulances to the exit side of the tunnel (outside of downtown) during such saturation conditions. Here the discrete time instants for possible changes in travel times may represent 'periodic location review times' associated with assessing the possibilities of ambulance relocation during rush hours. Model P2 however assumes unlimited service capacity, with service by the closest facility, thus limiting its applicability to those ambulance services having low ambulance utilizations (i.e., low probabilities of being unavailable). Model P2's assumptions are somewhat less restrictive when considering certain types of customer-to-facility-systems, including mobile outpatient clinics, medical testing services, lunch vans, book-mobiles, etc. The optimization problem with P2 is to find the optimal set of facility locations for each network state and the associated relocation rules that direct facilities when the network changes state. The cost function to be minimized is a weighted sum of the mean travel time to customers and of the cost of facility relocation.

P3: *Congested medians.* The formulation described here changes focus from probabilistic link lengths to probabilistic availability of servers (or facilities). It is the same as the Q-median problem with the complication that each facility (or server) may or may not be available to provide immediate service. Service is provided by the closest *available* server or, if all servers are busy, the customer is handled by some back-up service system. This model is most relevant to emergency service systems such as police, fire, emergency medical and emergency repair. The optimization objective is to find the set of Q facility locations that minimizes mean travel time associated with a random service request, where the average is computed over all possible combinations of servers available and unavailable. A special case of this formulation is the hypercube model [14], which has been utilized by several cities in North America and Europe.

P4: *Congested movable server systems.* This formulation is similar to P3, with the additional complexity that servers are movable and a server that is *en route* to a location may be assigned (dispatched) to a customer. This model combines all of the essential ingredients of actual ambulance systems except probabilistic travel times. The movable servers in this instance are allowed in order to be able to respond to temporary conditions of congestion in certain parts of the network; called *relocation* or *move-up* in fire services, the aim of relocating servers is to maintain in real time at least some minimal level of coverage in each area of the city (or network). The possible assignment of moving servers to customers assumes perfect knowledge of real-time server locations. To date, emphasis in P4 has been on determining a computerized server relocation policy for a *given* set of Q possible locations.

Throughout our paper emphasis is on model assumptions and results. Thus, no formal proofs will be given, but appropriate references are provided for the interested reader. Illustrative examples have also been included.

2. P2: Mobile servers on stochastic networks

Let $G(N, L)$ be an undirected network with N the set of nodes ($|N| = n$) and L the set of links. Service demands are generated exclusively at the nodes of $G(N, L)$ with h_i being the conditional probability that a demand comes from node i ($i \in N$) given that a demand was generated (h_i can be viewed as the 'normalized weight' of the node i).

In problem P2 travel times on the network change in a probabilistic manner according to the following simple model: At constantly spaced intervals (*epochs*) $G(N, L)$ is assumed to undergo changes of *state*. If r and s are two distinct states of the network and if $t_\beta(i, j)$ indicates the travel time on link $(i, j) \in L$ when the network is in state β, then $t_r(i, j) \neq t_s(i, j)$ for at least one link $(i, j) \in L$. Let M be the set of all possible states of G, $|M| = m$.

Transitions between network states at the epochs are governed by an ergodic Markov transition matrix P with $p_{rs} \in P$ being the probability of a transition from a state r to a state s ($r \in M$, $s \in M$). We also denote the steady–state probability vector of the matrix P as Π ($\Pi P = \Pi$, $\Sigma_{r=1}^{m} \Pi_r = 1$). The Q mobile servers (of infinite service capacity) which are to be located on the network are operated as follows: Whenever there is a demand for service, that customer is assigned by a system operator and *travels to the server* closest to it in terms of travel time. A customer not using such a system operator is assumed to know the current travel times and server locations, always selecting the closest server. Whenever there is a change of state of the network, the operator of the service has the option of *relocating* one or more of the servers. A relocation of a server is associated with a cost, which we shall choose to express in units of travel time. The operator's objective is to minimize the long-term expected cost (again expressed in terms of units of travel time) of providing the service. The long-term cost per epoch will be a weighted sum of the total expected travel time of demands to the servers per epoch (under all states of the network) and of the expected cost of the server relocations that take place per epoch.

We now define some additional quantities needed to express this problem in quantitative terms. Let $K(r) = \{k_1(r), k_2(r), ..., k_Q(r)\}$ be a set of Q points, where the Q servers are located when the network is in state r. We shall denote the shortest travel time between any two points x and y on G when the network is in state r as $d_r(x, y)$; the shortest travel time between any point in the set $K(r)$ and a specific point x on G when the network is in state s as $d_s(K(r), x)$; and the shortest travel time between the αth point in the set $K(r)$ and the γth point in the set $K(s)$ (for α and $\gamma = 1, 2, ..., Q$) when the network is in state l as $d_l(K_\alpha(r), K_\gamma(s))$.

The cost (in units of travel time) of relocating the αth server in $K(r)$ to the γth location in $K(s)$ with the network in state s is given by $f[d_s(K_\alpha(r), K_\gamma(s))]$. We also define binary variables $W_s(K_\alpha(r), K_\gamma(s))$ as follows: if the server at $K_\alpha(r)$ is relocated to the location $K_\gamma(s)$ when the state of the network changes from r to s, then $W_s(K_\alpha(r), K_\gamma(s)) = 1$; otherwise it is equal to 0.

Finally, we define as a *strategy*, any vector $K = (K(1), K(2), ..., K(m))$ with m elements, where each element $K(r)$, $r \in M$, provides the set of Q locations where the servers will be placed when the network is in state r. A *simple strategy* is any strategy with $K(1) = K(2) = \cdots = K(m)$, i.e., a strategy in which servers remain stationary under all states of the network.

We shall now state the assumptions under which the results for this model have been derived:

(1) The travel times $t_r(i, j)$ for all $r \in M$ and all $(i, j) \in L$ are finite.

(2) The time required to travel a fraction θ $(0 \leqslant \theta \leqslant 1)$ of any link $(i, j) \in L$ for any state $r \in M$ is equal to $\theta \cdot t_r(i, j)$.

(3) The current state of the network is known to the service operator (or, in the absence of an operator, to the customers) at all times.

(4) Time intervals between changes of state are much longer than trip times on the network.

(5) No demands or further state changes occur while servers are being relocated after a change of network state.

(6) All servers are available whenever a demand occurs. (This assumption is the same as that in the classic Q-median problem.)

(7) The relocation cost function, $f(\cdot)$, is non-decreasing concave.

(8) Demands are always assigned to the closest mobile server.

Assumption 1 assures connectivity of the network under all states. A less restrictive version of Assumption 1 (one that allows some of the $t_r(i, j)$ to be infinite while still leading to the same results) is given in [1,16]. Assumption 2 concerning uniformity of travel time on any given link is necessary in the proof of Theorem 1. Assumption 3 allows all travel in the network to be along the shortest path. Assumption 4 renders negligible the probability that link travel times will change while a customer is travelling to a server. (Were this to happen the server might no longer be travelling on a shortest travel time path.)

Assumptions 5 and 6 are the major simplifying assumptions in P2. Both assumptions would be approximately true, in practice, if, for instance, the average time interval between generation of demands on the network was much longer than travel times on the network (assuming that demands are generated according to a stationary renewal process independently of the state of the servers). As mentioned in the Introduction, Assumption 6 would be approximately true if each movable server in a customer-to-server system could simultaneously serve any number of customers (e.g., mobile libraries or 'bloodmobiles').

Assumption 7 is necessary for Theorem 1 to hold. It implies 'economies of scale' for the cost of travel times – a reasonable hypothesis in most practical contexts. (Obviously the family of acceptable functions, $f(\cdot)$, also includes the linear cost function.) Assumption 8, stating that the closest server to a customer always services that customer, is identical to that of the standard Q-median problem.

2.1. The problem

We can now express our objective function in terms of the quantities that have been defined. For any given strategy $K = (K(1), K(2), ..., K(m))$, the quantity

$$A = \sum_{r=1}^{m} \sum_{i=1}^{n} \pi_r h_i d_r(K(r), i) \qquad (1)$$

gives the long term ('steady–state') expected travel time per dispatch or server assignment. Similarly the quantity

$$B = c \sum_{r=1}^{m} \sum_{\substack{l=1 \\ l \neq r}}^{m} \Pi_r p_{rl} [\sum_{\alpha=1}^{Q} \sum_{\gamma=1}^{Q} W_l(K_\alpha(r), K_\gamma(l))$$

$$\times f[d_l(K_\alpha(r), K_\gamma(l))]] \qquad (2)$$

represents the long-term expected cost of server relocations per transition epoch, taking into account all possible changes of state from any possible state. The constant c in (2) is the relative weight assigned to the expected server relocation costs per epoch, in order to express relocation costs in units of travel time.

If an average of λ customers arrive per transition epoch, our problem is to minimize

$$Z = \lambda A + B . \qquad (3)$$

2.2. Main result

Our main results for problem P2 can now be summarized, beginning with the following Hakimi like

Theorem 1. *At least one set of optimal locations for P2 exists on the nodes of the network.*

The proof of Theorem 1, although long, is quite straightforward (see [1]). Basically it uses the properties of concave functions to show that Z cannot increase if a server which is located on some link $(i, j) \in L$ when G is in some state $r \in M$, is moved to one of the nodes i or j, no matter where the other $Q - 1$ servers are located. Theorem 1 reduces the total number of strategies to be considered to $\binom{n}{Q}^m$, since it reduces the number of candidate locations to $\binom{n}{Q}$ for each of the m states of the network. It also leads to a straightforward integer programming formulation of P2 [1]. The size of this integer programming problem grows very quickly as n, Q, and m increase. We have little computational experience with problems of this type to date. However, due to considerable similarity with the formulation of the deterministic median problem, the recent research of Erlenkotter [5], Garfinkel, Neebe and Rao [7], Revelle and Swain [18], Galvao [6], Jarvinen, Rajala and Sinervo [11], and Cornuejols, Fisher and Nemhauser [4] can perhaps be of help in solving our problem as well.

2.3. An example

The following simple example illustrates some of the ideas above. The network of Fig. 1 can be in one of the two states, 1 and 2. The numbers next to the links represent lengths (travel times), whereas those next to the nodes are the weights, h_i. Obviously, the only difference between the two states is in the travel time on link (3, 2). Customers appear at a rate of λ = 1 customer per transition epoch.

The Markovian transition matrix, P, that describes the statistical dependence between the two states is shown below.

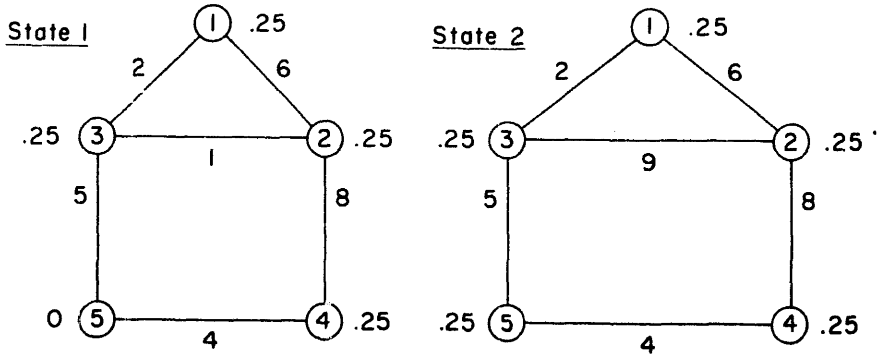

Fig. 1. The network under states 1 and 2.

$$P = \begin{array}{c|cc} \text{state}\diagdown\text{state} & 1 & 2 \\ \hline 1 & 0.25 & 0.75 \\ 2 & 0.5 & 0.5 \end{array}$$

The steady state probabilities associated with P are $\Pi_1 = 0.4$ and $\Pi_2 = 0.6$.

Let $f(y) = b\sqrt{y}$ be the relocation cost function, an increasing concave function (b is a positive constant) and let $c = 1$. Suppose that we wish to locate two servers on the network. It is easy to show that for $0 \leqslant b \leqslant 0.2357$ the optimal strategy is $\{K(1), K(2)\} = \{(3, 4), (1, 4)\}$, i.e., one of the servers must be relocated from node 3 to node 1 when the network makes a transition from state 1 to state 2 (and vice versa when the change of states is from 2 to 1), while the second server remains stationary at node 4 irrespective of the state of the network. Note that by moving from node 3 to node 1 (and back) as the network changes states, the first server retains some proximity to node 2.

On the other hand, when $b > 0.2357$ the optimal strategy is a simple one $\{K(1), K(2)\} = \{(1, 4), (1, 4)\}$. Note that, in this case, demands at node 2 are serviced by the server at node 1 when the state of the network is either 1 or 2 but not via the same shortest path.

2.4. *Additional results*

Ongoing research has led to some additional findings regarding P2 [3]:

(1) For the case in which a single mobile server is to be located on a probabilistic tree, the optimal strategy is to keep the server *stationary* at a single node, independently of the state of the tree. That node is the (single) median of the tree which remains unchanged under *all* states of the tree.

(2) For the case in which a single mobile server is to be located on a probabilistic network with m states, we have devised a heuristic algorithm which, in effect, amounts to solving a succession of classical 1-median problems, i.e. 1-median problems on deterministic networks. These latter problems can be solved efficiently and in a straightforward way (see, for instance, [8]).

(3) Simple upper and lower bounds on the minimum value of Z can be found for the most general case of Q mobile servers with m network states. By solving *m independent Q*-median problems, one for each network state, both an upper and a lower bound on Z can be obtained. A second upper bound is the value of Z associated with the best available *simple* strategy (this is also the solution to P1, a special case of P2).

3. P3: Congested medians

For P3 and for the remainder of this paper, travel times on the links of $G(N, L)$ are assumed deterministic. (This also permits some simplification in our notation.) The concern in P3 is with locating Q stationary facilities each having one server which can become busy (or unavailable). The objective is to locate the Q facilities so as to minimize the expected travel time associated with a random service request weighted appropriately by the equilibrium (steady–state) state probabilities of the system. In contrast to P2, 'states' here are defined according to the status (busy/available) of each of the facilities (and has nothing to do with travel times on the links).

3.1. *A one-server example*

To motivate our discussion of P3 and to demonstrate some general characteristics of spatially distributed queuing systems, we consider first a simple one-server example:

Suppose that an ambulance is to be garaged at a hospital located at point x on a roadway (link) connecting towns (nodes) a and b (Fig. 2). A fraction h_a (h_b) of ambulance demand is generated from town a (b). ($h_a + h_b = 1$.) No demands occur on the link. In response to each demand, or service request, the ambulance, if available, travels to the appropriate node at a speed v, spends a fixed time τ at the scene, then travels back to the hospital (with the patient on board), again at a speed v. If the ambulance is not immediately available, due to servicing an earlier demand, then the current service request is entered into a queue of other delayed service requests.

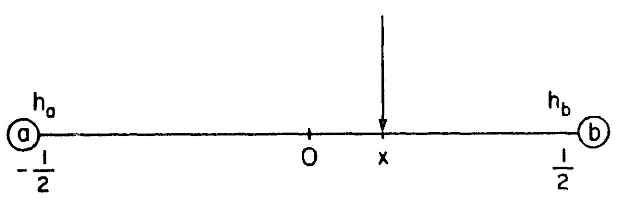

Fig. 2. Single link, two-node ambulance example.

Queued requests are serviced in a first-come, first-served manner. The *service-time, S*, associated with a service request is the sum of the travel time to the scene, the on-scene time, and the travel time back to the hospital.

The mean and variance of the service time are readily computed,

$$E[S] = \overline{S} = \tau + [1 + 2x(h_a - h_b)]/v\,, \tag{4a}$$

$$\mathrm{E}[(S - \overline{S})^2] = \sigma_s^2 = 4x^2[1 - (h_a - h_b)^2]/v^2 \tag{4b}$$

Now suppose that the total demand rate, summed over both towns, is λ ambulance requests per unit time. The ambulance *utilization factor,* or fraction of time the ambulance is busy, is

$$\rho = \lambda\overline{S} = \lambda\{\tau + [1 + 2x(h_a - h_b)]/v\}\,, \tag{5}$$

In order for an equilibrium (steady-state) solution to exist, we must have $\rho < 1$. Note that certain locations x, for given λ, v, h_a, and h_b, may violate this assumption, causing the queue to grow without bound; whereas other locations may satisfy $\rho < 1$.

In this example, we want to locate the hospital and ambulance at a point $x^* \in [-\frac{1}{2}, \frac{1}{2}]$ which minimizes the mean total *response time* (in-queue plus travel time) required for the ambulance to reach a random service request. If W_q is the mean in-queue time, then this mean response time can be written

$$\overline{T}_{\mathrm{R}} = W_q + \tfrac{1}{2}[1 + 2x(h_a - h_b)]/v\,. \tag{6}$$

Now to complete our specification of this locational problem, we need only to describe the nature of the stochastic process that generates service requests from a and b. This, in turn, determines W_a. Motivated by physical considerations, we assume that each town generates ambulance demands independently as a homogeneous Poisson process at a rate λh_a (town a) or λh_b (town b) per unit time. The completed model is a single-server queue having Poisson input and general independent service times, abbreviated M/G/1 in queuing parlance. For such a queue it is well known that [13]

$$W_q = \frac{\rho\overline{S} + \lambda\sigma_s^2}{2(1 - \rho)}\,. \tag{7}$$

Using (5) and (7) with (6), we can obtain an expression for the mean response time $\overline{T}_{\mathrm{R}}$ as a function of given system parameters and the locational decision variable x. Our objective is to select $x^* \in [-\frac{1}{2}, \frac{1}{2}]$ which minimizes (6).

It is a straightforward exercise in differential calculus to find the required minimum. The essential point is that for certain sets of values of system parameters the optimal location is at one of the nodes; for other sets of values, however, the optimal location is at some interior point $x \in (-\frac{1}{2}, \frac{1}{2})$. For instance, for a symmetric system in which $h_a = h_b = \frac{1}{2}$, the mean service time does not depend on x; thus ρ does not depend on x. The only quantity in (6) that depends on x is the quadratic in the numerator of W_q, $\sigma_s^2 = 4x^2/v^2$, which is clearly minimized at $x = 0$. Thus, assuming $\rho < 1$, for a symmetric system (i.e., $h_a = h_b$) of the type we are discussing, the optimal facility location is always at the link mid-point, $x = 0$. Moreover, this optimal location moves away from $x = 0$ continuously as $(h_a - h_b)$ moves continuously from 0.

To illustrate typical behavior, consider an example for which $\tau = 1$, $v = 1$, $\lambda = \frac{1}{4}$. If the ambulance were located at $x = 0$, we would have

$$\overline{S} = \tau + 1/1 = 1 + 1 = 2; \quad \sigma_s^2 = 0\,; \; \rho = \tfrac{1}{4}\cdot 2 = \tfrac{1}{2}\,.$$

(Note that one-half of the mean service time in this instance is on-scene time and the other half is travel time). Regardless of h_a and h_b, an ambulance located at $x = 0$ would provide the following service performance: mean travel time to the scene $= \frac{1}{2}$; $W_q = [\frac{1}{2}\cdot 2]/2(\frac{1}{2}) = 1$; $\overline{T}_{\mathrm{R}} = 1 + \frac{1}{2} = \frac{3}{2}$. Thus, any movement away from $x = 0$ (for $h_a \neq h_b$) must reduce $\overline{T}_{\mathrm{R}}$ below $\frac{3}{2}$. Such movement would increase σ_s^2 above 0, but may decrease S and ρ, resulting in a net decrease of the objective function $\overline{T}_{\mathrm{R}}$. The optimal location x^* is shown as a function of $(h_b - h_a)$ in Fig. 3. Note that for $|h_b - h_a| \gtrsim 0.3$, the optimal location is at the busier of the two nodes; for all other values of $(h_b - h_a)$, x^* is an interior point. In fact, if no queuing ever occurred, the objective function to emerge is simply $\frac{1}{2}[1 + 2x(h_a - h_b)]/v$, a linear function whose minimum exists at the busier of the two nodes a or b (or if $h_a = h_b$, at any point $x \in [-\frac{1}{2}, \frac{1}{2}]$). This result also obtains if a queue is *not allowed to form,* by assigning customers arriving when the server is busy to a back-up system, where 'cost' per response is assumed higher than that of the primary system.

The observations from the M/G/1 analysis have motivated our general treatment of problem P3, having multiple-servers. Our results to date include only a class of subproblems for which nodal solutions are found to be optimal. However, the reader must not think that general spatially distributed queuing

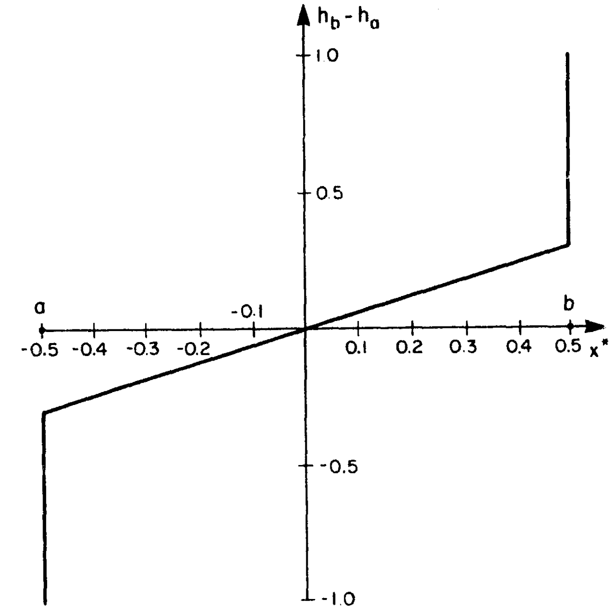

Fig. 3. M/G/1 example: optimal location of facility as a function of $h_b - h_a$.

systems will always exhibit locational optimality at nodes.

3.2. General formulation

Let $\mathbf{X}_Q$ be the set of all possible locations of Q facilities ($Q > 1$), on the network G, i.e.,

$$\mathbf{X}_Q = \{X_Q = (x_1, ..., x_Q); x_k \in G, k = 1, ..., Q\} .$$

Given any location $X_Q = (x_1, ..., x_Q) \in \mathbf{X}_Q$, let $\hat{x}_k$ denote that the facility at x_k is not staffed with an available server (the facility is busy) and $\hat{\hat{x}}_k$ that the facility x_k does have an available server. Therefore, for any $X_Q \in \mathbf{X}_Q$ there are 2^Q combinations (states), any one of which the network can assume, according to the status of the Q facilities. Let $Y_{X(Q)}$ be the set of all states for $X_Q \in \mathbf{X}_Q$ and let $y_{X(Q)}$ (or for convenience y_Q) be a generic element of $Y_{X(Q)}$.

Let $t(i, j)$ be the travel time on link (i, j), $(i, j) \in L$, and let $d(y_Q, j)$ be the (minimum) travel time associated with the closest *available* server [1] to node j, when the system is in state y_Q.

Our results for P3 have been obtained under the following set of assumptions:

(1) Demands (service requests) occur according to a general renewal process.

(2) Each request requires a total service time (=on-scene service time plus travel time) whose distribution is general and not dependent on the identity of the server or the history of the system. This implies that on-scene service times are assumed to be much greater than travel times so that variations in total service times which are due solely to variations in travel times among potential servers can be ignored.

(3) Appropriate ergodicity conditions apply so that a unique steady state distribution exists.

(4) The time required to travel a fraction θ ($0 \leqslant \theta \leqslant 1$) of any link $(i, j) \in L$ is equal to $\theta \cdot t(i, j)$.

Assumption 2 is reasonable for many urban service systems which are often characterized by on-scene service roughly an order of magnitude greater than typical travel times. Assumption 4 is used in the proof of Theorem 2 below.

For any possible set of locations $X_Q \in \mathbf{X}_Q$, let $P(y_Q)$ be the steady-state probability that the network is in state $Y_X \in Y_{X(Q)}$. Let y_Q^0 be the state in which all the Q facilities are busy (i.e., $y_Q^0 = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_Q)$ in our notation).

Conditioned on any state $y_Q \in Y_{X(Q)} - \{y_Q^0\}$, the expression

$$\sum_{j=1}^{n} h_j d(y_Q, j)$$

is the expected travel time associated with a random service request. If, however, the system is in state y_Q^0 (i.e., all Q servers are busy), when a service request occurs, we will assume that the request is handled by a back-up system having a travel time cost R per response (R assumed larger than any travel time in the primary system). (We have also obtained results for systems that allow queuing, under specific queuing service disciplines and assuming negative exponential service times.)

The congested median problem is now stated:

$$\min_{X_Q \in \mathbf{X}_Q} \quad F(X_Q) \tag{8}$$

[1] Such a zero-lookahead strategy is very reasonable, but not always optimal in the sense of minimizing time-average mean travel time. An optimal policy occasionally requires assignment of other than the closest available server [12], in order to leave the system in a state which best anticipates future service requests. We do not consider such strategies in our formulation of the congested median problem.

with

$$F(X_Q) = \sum_{y_Q \in Y_{X(Q)} - \{y_Q^0\}} P(y_Q) \sum_{j=1}^{n} h_j d(y_Q, j)$$

$$+ P(y_Q^0) \cdot R \, .$$

Obviously, the classic median problem is a special case of (8) arising when $P(y_Q) = 0$, $\forall y_Q \neq (\hat{x}_1, ..., \hat{x}_Q)$ – the state where all the units are available. The weights $P(y_Q)$ in (8) represent the fraction of time that the network is in each of the 2^Q possible states. Therefore, as noted before, we take into account the fact that any subset of facilities can become depleted of servers.

3.3. Results

Our main results for P3 begin once more with a Hakimi-like theorem:

Theorem 2. *At least one optimal solution to* (8) *exists on the nodes of the network.*

The proof of Theorem 2 is given in [2]. We believe that it can be extended to more general dispatch policies, using criteria other than 'closest-available server' to assign demands to servers.

Once the solution space has been limited to the nodes of G, the following important observation is useful in solving the congested median problem for specific situations: Whenever

(a) demands over the nodes of G are generated in a Poisson manner at a total rate λ and at each node j independently in a Poisson manner with rate $\lambda_j = h_j \lambda$,

(b) the total service time for each server is negative exponential with mean $1/\mu$, and

(c) each facility can hold exactly one server, then (8) can be viewed as a locational objective function for the hypercube queuing model [14], (see [15] for an overview of the hypercube model).

In [12] Jarvis developed a heuristic algorithm for finding a set of, hopefully, near optimum locations in the framework of the hypercube model where locations are constrained to nodes and each node can contain not more than one facility.

For the purposes of P3, it should be noted that, because of Theorem 2, both the hypercube model and Jarvis' algorithm do not suffer from a loss of generality by considering locations only on nodes. We have applied Jarvis' algorithm, in the manner described by Larson [15], to several moderate-size network problems (e.g., $n = 25$, $Q = 4$) using (8) as the performance measure to be optimized. Convergence of the algorithm has usually been fast and the results have been surprising on occasion. It is not uncommon to find a solution consisting of a set of locations which is among the weakest possible sets in terms of the classic (deterministic) median problem.

3.4. Example

The following example will illustrate some of our previous discussion. Suppose we want to locate three facilities on the simple network shown in Fig. 4. The numbers next to the nodes are the fractions of demands from each node j, $j = 1, ..., 5$ and the numbers next to the links are the travel times. There are $\binom{5}{2}$ possible distinct locations:

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5},

{1, 4, 5}, {2, 3, 4}, {2, 3, 5}, {2, 4, 5}, {3, 4, 5}.

The optimal location according to the standard 3-median problem is {1, 2, 5}, which can be obtained by hand computation. Suppose however that service requests occur in the network in a Poisson fashion with $\lambda = 4$, and the service time for each one of the three units is exponential with identical means $\mu^{-1} = 1$. Let us assume a zero capacity queue with $R = 5$ units of time – the cost resulting when dispatching the reserve unit. We also assume that server preferences are determined solely by geographical proximity.

The Jarvis algorithm with an initial location at the absolute 3-median, i.e., {1, 2, 5}, converges after one iteration to the optimal solution at location {2, 3, 5}. The improvement achieved by moving from the location {1, 2, 5} to {2, 3, 5} is 3% in terms of the objective function of the congested median problem. It is interesting to realize that the location {2, 3, 5} is among the weakest possible locations in terms of the

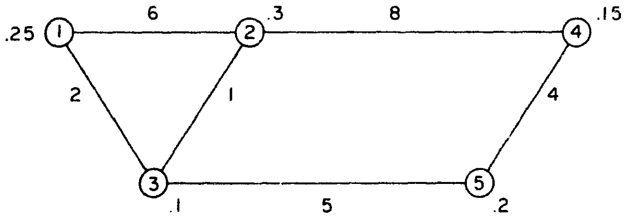

Fig. 4. A simple 5 node network.

standard median problem. This indicates that blind application of the classic (deterministic) median problem can lead to erroneous results, even for such simple networks.

4. P4: Congested systems with mobile servers

This problem was the main topic of research of one of the authors' doctoral dissertation (see [1, Chapters 3, 4, and 6]). It is also the problem (among P1–P4) with which our computational experience is most extensive. Due to space limitations our discussion of P4 will be limited and mostly descriptive.

In P4 our concern is not so much with initially locating Q mobile servers but in exploring how free servers at any given time should be *repositioned* to 'compensate' for servers which are busy and, therefore, unavailable. The objective is to minimize a weighted sum of the expected travel time to a random request for service and the expected server repositioning costs per unit of time. It is always assumed that each request for service is serviced by the *closest available* server, with the exception of requests which find all servers busy. These latter requests are presumed lost and a penalty, R, is paid whenever such an event occurs in a manner analogous to that described earlier for problem P3. A related assumption is that an idle server is eligible for being assigned to a request for service, *even* while in the process of moving, from one location to another.

In order to provide a proper focus for indicating the type of approach taken as well as the limitations of our research, let us consider a case in which $Q = 3$. Assume that somehow it has been determined that the best set of locations for the three (indistinguishable) servers, whenever all three are available, is at the nodes $\{i, j, k\}$ and that at a particular moment the three-server system has *just entered* state $(\hat{i}, \hat{\hat{j}}, \hat{k})$ as described by the status of each of the three servers, in the notation described under P3. If servers are restricted to locations i, j, and k only (these, for instance, might be the locations of fire houses) then the following three actions are possible when the system enters state $(\hat{i}, \hat{\hat{j}}, \hat{k})$:

(i) do nothing;

(ii) move an idle server from node j to node i;

(iii) move an idle server from node k to node i.

If action (i) is taken, the possible successor states to $(\hat{i}, \hat{\hat{j}}, \hat{k})$ are $(\hat{\hat{i}}, \hat{\hat{j}}, \hat{k})$ if the next event to take place is the completion of service to the call with which the server from i is currently occupied and $(\hat{i}, \hat{j}, \hat{\hat{k}})$ or $(\hat{i}, \hat{\hat{j}}, \hat{k})$ if the next event is a request for service closet to j or to k, respectively. The immediate cost associated with 'do nothing' is the sum of the expected travel times associated with servicing demands from j or from k (as the case may be) weighted by the probabilities of transitions to $(\hat{i}, \hat{j}, \hat{\hat{k}})$ or to $(\hat{i}, \hat{\hat{j}}, \hat{k})$ respectively.

If action (ii) is taken the possible successor states are: $(\hat{\hat{i}}, \hat{\hat{j}}, \hat{\hat{k}})$, if the next event is a completion of the busy unit's service; $(\hat{i}, \hat{j}, \hat{\hat{k}})$ if the next event is a request for service *and* the server which was moving from j to i happened (in mid-trip) to be closer than the available server at k to the request for service, at the instant when the request for service occurred; $(\hat{\hat{i}}, \hat{j}, \hat{k})$ if the next event is a request for service *and* the server at k is closer to the request for service than the moving (from j to i) server at the instant when the request for service occurs. An important assumption is implicit in the statement above: during the repositioning period only one completion of service or arrival of a new service request (but not both) is possible. A direct consequence is that if a service request occurs during the repositioning period, and if the moving server is not dispatched to it, then the next service request can occur only after the repositioned server reaches its destination. This assumption facilitates the analysis that focuses on time instants following a transition. The expected immediate cost for action (ii) takes into consideration the possibility that the moving server will be dispatched to the next service request. The situation for action (iii) is entirely symmetrical to that for action (ii).

The last two paragraphs discussed transitions, transition probabilities and immediate costs associated with state $(\hat{i}, \hat{\hat{j}}, \hat{k})$. A similar analysis can be performed for all seven other states. Note that for states $(\hat{\hat{i}}, \hat{\hat{j}}, \hat{\hat{k}})$ and $(\hat{i}, \hat{j}, \hat{k})$ only the 'do nothing' action is available.

We are now in a position to present our assumptions and outline the results that we have obtained for P4:

(1) Demands over the nodes of G are assumed generated in a Poisson manner at a total rate λ and at each node j independently in a Poisson manner with rate $\lambda_j = h_j\lambda$.

(2) The total service time for each server is negative exponential with mean $1/\mu$.

(3) $1/\mu >>$ (travel times on the network) so that all variations in service times due to the travel times can be ignored. As noted under P3 this assumption is

reasonable for several important urban services.

(4) All servers travel at the same constant speed on the network G.

(5) Only one server can be repositioned at any one time. This implies that, for $Q = 4$, when two servers are busy and the other two idle, it is not permitted to reposition simultaneously the two idle servers to the positions of the two busy servers. (This, however, does not preclude repositioning first the one server and then the other.)

(6) Perfect information about the position of any moving (due to repositioning) or stationary servers is available at all times to the server dispatcher.

Assumption 6 is reasonable in view of the availability of increasingly accurate and reliable 'automatic vehicle location' (AVL) systems for urban service systems. Unfortunately, the only reason for assumptions 4 and 5 is that we have found the analysis to be mathematically intractable without them.

4.1. Results

As the discussion of states, state-transition probabilities, alternative actions and immediate costs may have suggested, we have been able to approach and solve the version of P4 described above as a Markovian decision process (MDP). For any *given* set of Q nodal locations $\{i_1, i_2, ..., i_Q\}$ housing the Q servers (the 'home locations'), this solution finds the optimum repositioning policy that minimizes a weighted sum of the expected travel time to a random request for service and the expected server repositioning cost per unit of time. This objective function is for MDP's the 'average cost per unit of time', usually denoted as g. An optimum policy is a listing of the actions to take for each possible state of the system in order to minimize g.

A rather involved computer program (about 900 APL statements) has been written to accomplish this. The program calculates all transition probabilities and immediate costs for all possible states and permissible actions and then applies Howard's policy iteration algorithm [10] to find the optimum policy and the associated minimum value of g. We have solved many problems of moderate size ($n = 25$, $Q = 3$) in order to explore the sensitivity of the optimum policy and of g to the server utilization ratio $\rho(=\lambda/Q\mu)$ and the "goodness" of the home locations [1]. Reductions of up to 20% in the objective function, g, over the 'all-do-nothing' policy (i.e., no repositioning allowed) have been observed in the examples attempted. The greatest benefits from repositioning have been observed at the middle-range of server utilization rates, which is also the most interesting and common range for the applications with which we are concerned. The reasons for this are quite obvious: at very low server utilization rates, dispatching to requests for service is so infrequent that the benefits resulting from reduced travel times due to server repositioning are too small to justify the effort of repositioning; by contrast, at very high utilization rates, the servers are busy most of the time, leaving little room for maneuverability by the operator of the service. In our numerical examples we have also found cases in which the optimum policy is unexpected and "counter-intuitive" (see Section 4.2).

An interesting by-product of this research has been the development of an efficient algorithm for keeping track of which available server is closest to each node of G as an available server moves (due to repositioning) between any two nodes on G.

We have found that the analysis of P4 with $Q = 2$ is relatively simple and that some special simplifying results can be derived for this case. In fact, for $Q = 2$, we have been able to develop results for the case in which the expected service times are different for each of the two-servers and for the case in which, when one of the two servers is busy, the other one can be repositioned to *any* node on G and not only to the other available home location.

4.2. Example

In this section we present an example intended to demonstrate that for high utilization factors and/or for large travel times along the repositioning paths, repositioning often occurs in anticipation that the moving unit will be near 'strong' nodes while travelling and will be dispatched to the next service request. To show that let us refer to the network in Fig. 5.

Let us consider the case where the two existing service units are located at nodes 1 and 24 and $R = 40$. It is easy to verify that node 24 is 'stronger' than node 1 ($\Sigma_{j=1}^{25} h_j\, d(24,j) = 15.84 < \Sigma_{j=1}^{25} h_j\, d(1,j) = 19.10$).

The model was applied for several cases when the utilization factor ρ varies from zero to infinity. As expected, for low utilizations the optimal policy is to move the service unit from node 1 to 24 when the network enters state $(\hat{1}, \hat{\hat{24}})$ and to do nothing

[2] Since we are dealing with a zero-line capacity queue, $\rho = \lambda/Q\mu$ is *not* the average function of time servers are busy. This latter quantity is less than ρ due to lost calls.

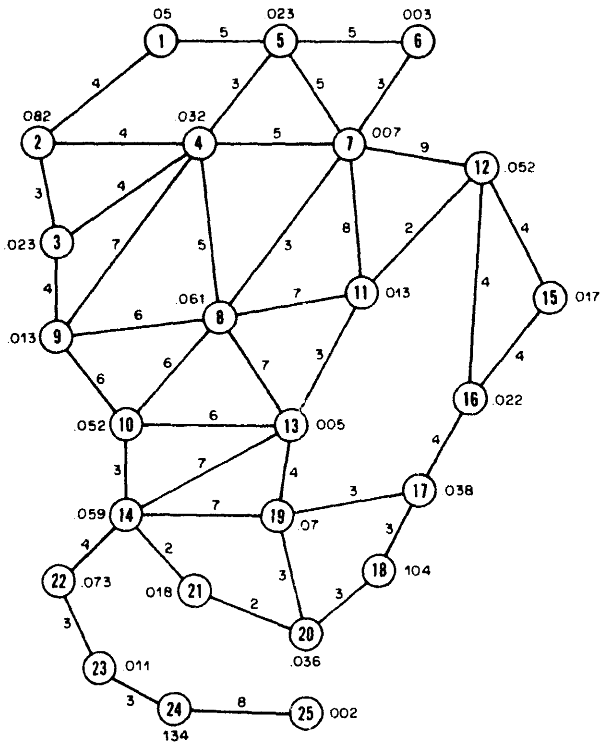

Fig. 5. An urban transportation network.

when the network enters state $(\hat{\hat{1}}, \hat{24})$. For high utilization, however, the optimal policy is to reposition the available server for *both* states $(\hat{1}, \hat{\hat{24}})$ and $(\hat{\hat{1}}, \hat{24})$. To understand this result, note first that the distance between those two nodes is quite large – 30 units of time. In addition to that, all the nodes within a distance of less than 26 units from node 24 are 'stronger' than node 24. In fact, the shortest path form node 24 to node 1 includes the two strongest node: node 10 and node 14 (which is also the absolute median). Therefore if the decision is to reposition for state $(\hat{1}, \hat{\hat{24}})$, then unless ρ is very small, chances are high that the repositioned server will be dispatched from a 'stronger' position than node 24. Only for very small utilizations is it likely that the repositioned server from node 24 might be dispatched near or at node 1, in spite of the large distance between nodes 1 and 24.

5. Conclusions

We have reviewed in this paper a class of problems which begin to bridge the gap between deterministic location formulations as represented by the median problem and an important area of urban applications characterized by probabilistic demands and travel times, facility unavailabilities (congestion) and server mobility. Several potentially useful results have been obtained under sometimes restrictive sets of assumptions. By examining the entries in Table 1, representing the types of complexities we have added to the standard Q-median formulations one can readily see that more work is needed in integrating the types of methods used in this paper. For instance, a model incorporating both probabilistic travel times and probabilistic availabilities of servers would be useful. Even for the complexities analyzed here, more work is needed on reducing the restrictiveness of certain assumptions and on devising computationally efficient algorithms. Finally, future work should also begin to include the all-too-important multi-objective nature of many urban service systems.

We are hopeful that the continued analysis of these problems will tie more closely together previously disparate efforts on network analysis, queuing theory, spatial analysis, and Markovian decision processes.

References

[1] O. Berman, Dynamic repositioning of mobile servers on networks, Technical Report No. 144, M.I.T. Operations Research Center, Cambridge, MA (1978).

[2] O. Berman and R.C. Larson, The congested median problem, Working Paper OR 076-78, M.I.T. Operations Research Center, Cambridge, MA (1978).

[3] O. Berman and A.R. Odoni, Locating mobile servers on a network with Markovian properties, Working Paper OR 083-78, M.I.T. Operations Research Center, Cambridge, MA (1978); revised version.

[4] G. Cornuejols, M.L. Fisher and G.L. Nemhauser, Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms, Management Sci. 23 (1977) 789–810.

[5] D. Erlenkotter, Dual-based procedure for uncapacitated facility location, Working Paper 261, Western Management Science Institute, University of California, Los Angeles, CA (1976).

[6] R.D. Galvao, A dual-bounded algorithm for the P-median problem, Paper presented at the International Symposium on Locational Decisions at Banff, Alberta, April 24–28, 1978.

[7] R.S. Garfinkel, A.W. Neebe and M.R. Rao, An algorithm for the m-median plant location problem, Transportation Sci. 8 (1974) 217–236.

[8] G.Y. Handler and P.B. Mirchandani, Location on Networks (M.I.T. Press, Cambridge, MA, 1979).

[9] S.L. Hakimi, Optimum locations of switching centers and the absolute centers and medians of a graph, Operations Res. 13 (1964) 462–475.

[10] R.A. Howard, Dynamic Programming and Markov Processes (M.I.T. Press, Cambridge, MA, 1960).

[11] P. Jarvinen, I. Rajala and H. Sinervo, A branch-and-bound algorithm for seeking the *p*-median, Operations Res. 20 (1972) 173–178.

[12] J.P. Jarvis, Optimization in stochastic service systems with distinguishable servers, Technical Report IRP-TR-19-75, M.I.T. Operations Research Center, Cambridge, MA (1975).

[13] L. Kieinrock, Queuing systems, Vol. 1: Theory (Wiley, New York, 1975) Chapter 5.

[14] R.C. Larson, A hypercube-queuing model for facility location and redistricting in urban emergency services, Comput. Operations Res. 1 (1974) 67–95.

[15] R.C. Larson, Structural system models for locational decisions: an example using the hypercube queuing model, in K.B. Haley, Ed., OR'78 (North-Holland, Amsterdam, 1979) 1054–1091.

[16] P.B. Mirchandani, Analysis of stochastic networks in emergency service systems, Technical Report IRP-TR-15-75, M.I.T. Operations Research Center, Cambridge, MA (1975).

[17] P.B. Mirchandani and A.R. Odoni, Locations of medians on stochastic networks, Transportation Sci. 13 (1979) 86–97.

[18] C. ReVelle and R.W. Swain, Central facilities location, Geograph. Anal. 2 (1970) 30–42.